First steps of the project

Task 2: Business Understanding Report

Project title: ERR News Content Analysis
Members of the team:

- Mihkel Rump
- August Roosi
- Andreas Närep

Repository:
https://github.com/August-Roosi/ERR-Uuring

Identifying your Business Goals

Background:

Conducting ERR content assessment from 2007 through in-depth analysis that shows how frequently people, organizations or places are mentioned and how article header and body length has changed over time. Moreover, the project will look at language usage changes in ERR News articles.

Business Goals:

Temporal Mention Analysis: Trends in mentions of individuals, organizations and places in ERR News articles between 2007 and now.
Language Evolution Analysis: Shifting use of language including changing common words and phrases across time.
Article Body and Header Analysis: Mapping out the length of the article body and header, showing how it has changed in accordance with time.

Business Success Criteria:

Temporal Mention Analysis:
Compile ERR News content for 2007 – present into a structured JSON file successfully.
Show trends about mentions of main entities over different time lengths

Language Change Analysis:
Prepare a comprehensive dataset on language variations over time.

Visualize the development of language with charts and graphs.

Article Body and Header Analysis:
Map and visualize the changes in article body and header lengths.

Assessing Your Situation:

Inventory of Resources:

- Data Sources: ERR News archives (2007-present), NLP tools, data storage infrastructure.
- Personnel: Data analysts, NLP special
- Technology: NLP tools, data processing scripts, visualization tools.

Requirements, Assumptions, and Constraints:

- Data Availability: Assumption that ERR News archives are available since 2007.
- NLP Tools Performance: Requirement for efficient NLP tools to extract meaningful insights.
- Project Timeline: Constraint to complete data gathering and analysis in a week.

Risks and Contingencies:

- Data Accessibility Issues: Risk of incomplete archives. Contingency: Develop a robust data validation process.
- NLP Tool Limitations: Risk of insufficient insights. Contingency: Explore alternative tools and techniques.

Terminology:

- Temporal Mention Analysis: Examination of how often specific entities are mentioned over time.
- Language Evolution Analysis: Study of changes in language patterns and word usage over different periods.
- Selenium: A powerful tool commonly used for automating web applications. In this project, Selenium is utilized to extract ERR News articles from the web, streamlining the data collection process.
- JSON: A lightweight data-interchange format that is easy for humans to read and write. It is also easy for machines to parse and generate. In the context of the ERR News project, JSON is used as a structured data format for storing and exchanging news data.
- NLP: Short for natural language processing refers to machine learning algorithms that are able to interpret and comprehend the meaning of words and sentences in human languages.

Costs and Benefits:

Costs: Investment in NLP tools, data storage, and potential challenges in data preprocessing?
Benefits: Discover historical trends in content? Know audience interests better? Improve the future content?

Defining Your Data-Mining Goals

Data-Mining Goals:

Data Compilation: Gather ERR News content from 2007 to the present. Use python library Selenium to dinamically crawl **https://www.err.ee/uudised.** Structure gathered data into a JSON file.
Entity Mention Analysis: Develop algorithms to identify and quantify mentions of individuals, organizations, and places.
Language Evolution Analysis: Utilize NLP techniques to analyze changes in language patterns over different time intervals.
Article Body and Header Analysis: Create methods that map out the lengths of article header and boy lengths making it into a presentation.

Data-Mining Success Criteria:

Successful Compilation:

- Compile ERR News content into a well-organized JSON file.
- Ensure data integrity and completeness.

Entity Mention Analysis:

- Refine the algorithm to conduct a precise Entity Mention Analysis, specifically focusing on accurately identifying and quantifying references to individuals, organizations, and places within the ERR News content.
- Present visualizations illustrating trends in entity mentions over time.

Language Evolution Analysis:

- Extract meaningful insights into language evolution.
- Provide visualizations and reports demonstrating language changes.

Article Body and Header Analysis:

- Provide graphs representing found data.

Gathering data:

Task 3: Data Understanding

Gathering data:

Articles and the titles of these articles are needed. The required data is available. It will be acquired with a web scraper from ERR (Eesti Rahvusringhääling). Addressing alternative data sources is not necessary since it is already confirmed that data can be mined from ERR. All of the articles on any given day in any given year are chosen. The timeframe from which the articles will be acquired is from about 2007 until 2023. As of now the data from the whole timeframe has not been gathered but just a part of it as it is a long process to scrape such a big amount of data and might have to be split into separate timeframes for convenience. But this shows that there aren't any issues gathering the data and everything works as expected.

Describing data:

As said in the first section, the data is acquired from ERR (Eesti Rahvusringhääling) by web scraping news articles from 2007 until 2023. All of the data is in a JSON file where the articles and the titles are categorized by their year. Every year contains a json key for every month and every month a key for every day in that month. And every day contains a json array where there are json objects for every article. Those articles consist of key-value pairs. There are two key-value pairs for every article, one for the header and one for the body of the article.

Exploring data:

Since the data hasn't been acquired fully, it can not be analyzed and examined thoroughly for this point. The data exploration was done on the small sample which has already been mined. There were no problems with the titles of articles but some of the bodies are empty, which need to be removed from the data. Also some of the articles which have been mined aren't actually articles, the body for these kinds of "articles" just state a show host. Here is an example:

```
{
        "päis": "ETV spordi lühiuudised, 22. november",
        "sisu": "Saatejuht Tarmo Tiisler\n \n"
}
```

These are the main faults in the data, which need to be eliminated.

Verifying data quality:

There aren’t any major concerns or problems with this data and it is usable for all of our goals that we want to achieve with our project. The biggest problems that occur right now is the absence of some of the article body’s and unnecessary bodies as shown in the previous section.

Task 4: Project plan

1. Build a webscraper to get some decades of articles written by ERR.(August Roosi, 5h)
    a. Use Selenium such that all the data is written into a json file or files.
    b. As there is no API we need to make Selenium click on buttons like person would, making the process somewhat slower.
2. Quantify and analayze the lengths of articles headers and contents, so that we get an overview of how the size of articles has changed over the years.(August Roosi, 3h)
3. Count the mentions of names, organitsations and places and see what err mentions the most.(Mihkel Rump, 8h)
4. Observe the usage of words over time and see how it has changed.(Andreas Närep, 8h)
    a. We also use EstNLTK here.
5. Make the poster adding the findings that we got. (Andreas Närep, August Roosi, Mihkel Rump, 2h)